Abstract

Web mining is a one class of data mining. Web Mining is a variation of data mining that distills untapped source of abundantly available free textual information. The need and importance of web mining is growing along with the massive volumes of data generated in web day-to-day life. Web data Clustering is the organization of a collection of web documents into clusters based on similarity. A good clustering algorithm should have high intra-cluster similarity and low inter-cluster similarity. The process of grouping similar documents for versatile applications has put the eye of researchers in this area. In general, web data always arrives in a continuous, multiple, rapid and time varying flow. The Researchers in web mining proposed many methods to extract web contents, but they are fail to handle dynamic data. Web content extraction algorithms are important to extract useful contents from web sources. We propose a new method for web content extraction. It consist of four phases: Web document selection phase, web cube creation phase, web document preprocessing phase, and presentation phase. In the first phase list of web documents are selected for mining, second phase documents are used to create web cube, third phase documents are preprocessed, in the final phase results are presented to users. The experimental results of proposed system are compared with existing
methods, Performance of proposed system is better than previous methods.

References

- Tasawar Hussain, Dr. Sohail Asghar and Nayyer Masood, "Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence", 6th International Conference on Emerging Technologies (ICET) IEEE, pp. 21-26, 2010.
- Shian-Hua Lin and Jan-Ming Ho. Discovering Informative Content Blocks from Web Documents, KDD-02, 2002.

Index Terms

Computer Science
Information Sciences

Keywords

Web Cube creation  Maintenance  Web document Cleaning  Web Mining